AI AGE

# Nvidia's GTC will mark an AI chip pivot. Here's why the CPU is taking center stage

PUBLISHED FRI, MAR 13 2026•3:00 PM EDT

Katie Tarasov
@KATIETARASOV

WATCH LIVE

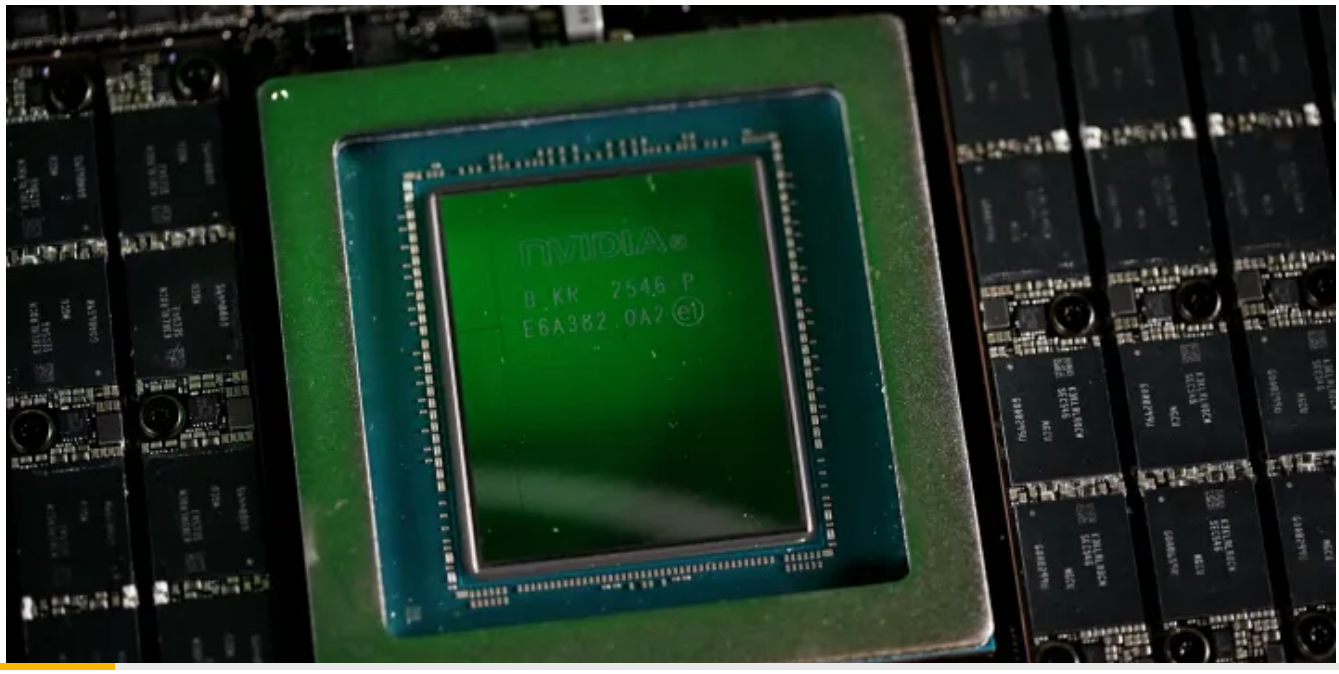
## KEY POINTS

- Central processing units (CPUs) are having a renaissance, and Nvidia says they're "becoming the bottleneck" amid surging agentic AI workflows.
- Intel and AMD remain the leaders in data center CPUs, but Nvidia is set to unveil new CPU details at its annual GTC AI conference next week.
- Nvidia shifted its CPU strategy in February with an announcement that standalone processors are now deployed in Meta data centers.

Follow your favorite stocks **CREATE FREE ACCOUNT**

**In this article**

NVDA -0.05 (-0.03%)




CNBC

**Nvidia showed CNBC its latest Vera CPU at its Santa Clara, California, headquarters on Feb. 13, 2026.**

*Marc Ganley | CNBC*

Nvidia's graphics processing units have been the hottest-selling chips for years, but the sudden advent of agentic artificial intelligence has brought on a renaissance for its more modest host chip, the central processing unit.

Now, Nvidia is poised to unveil new details about its agentic-optimized CPUs at its annual GTC conference that kicks off on Monday, with a CPU-only rack

likely to appear on the showroom floor.

"CPUs are becoming the bottleneck in terms of growing out this AI and agentic workflow," Dion Harris, Nvidia's head of AI infrastructure, told CNBC this week, calling it an "exciting opportunity."

The chip giant announced its first data center CPU, Grace, in 2021, and the next generation, Vera, is now in production. The CPUs are typically deployed alongside Nvidia's famous Hopper, Blackwell or Rubin GPUs in full rack-scale systems.

Exploding demand for GPUs has turned Nvidia into a household name and the most valuable publicly traded company in the world, with a $4.4 trillion market cap. Its broader chip strategy took a major turn in February, when Nvidia struck a multiyear deal with Meta ⊞ that included the first large-scale deployment of Grace CPUs on their own, with plans to deploy Vera in 2027.

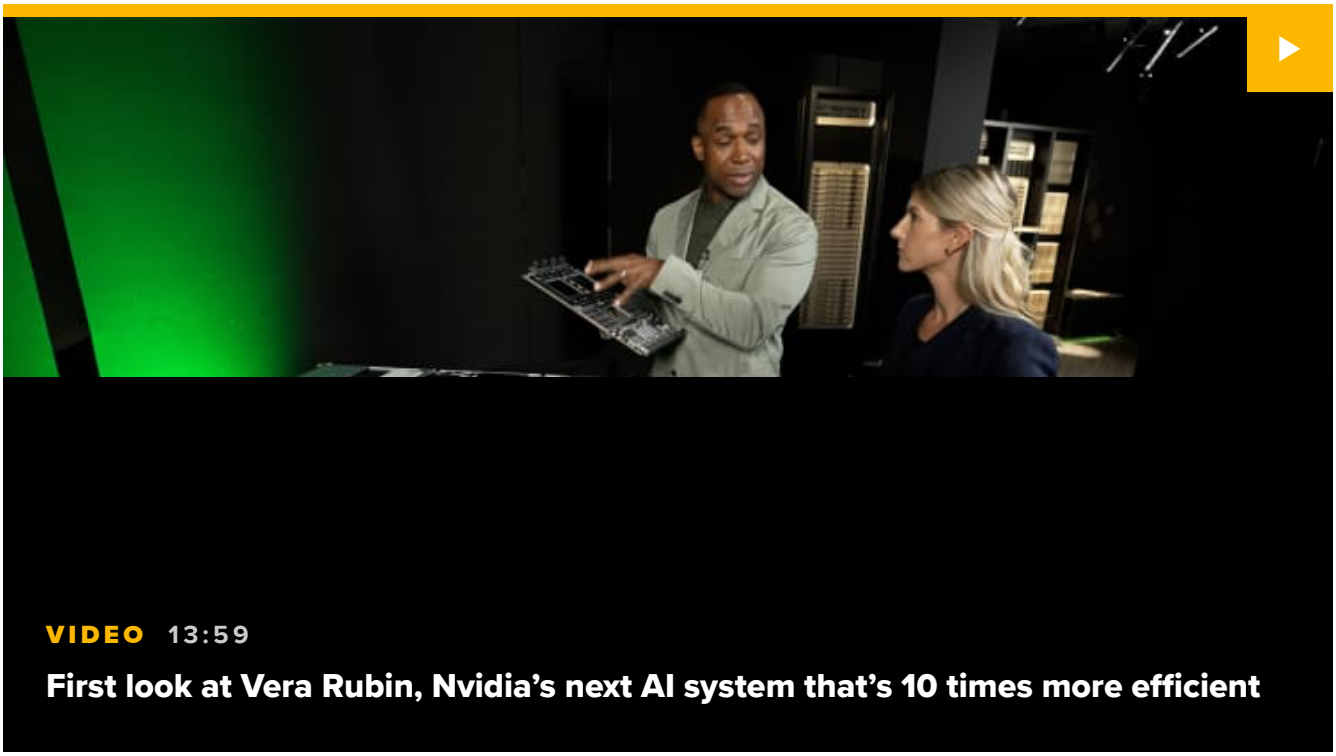
Thousands of standalone Nvidia CPUs are also helping power supercomputers at the Texas Advanced Computing Center and Los Alamos National Lab, Nvidia told CNBC.

Bank of America ⊞ predicts the CPU market could more than double, from $27 billion in 2025 to $60 billion by 2030. In the latest quarter alone, Nvidia generated data center revenue of over $62 billion, up 75% from a year earlier.

The CPU resurgence is driven by a fundamental change in compute needs, as mass AI adoption shifts from call-and-answer chatbots to task-oriented agentic apps.

While GPUs are ideal for training and running AI models because they have thousands of tiny cores narrowly focused on performing many operations simultaneously, CPUs have a smaller number of powerful cores running sequential general-purpose tasks.

Agentic AI requires a lot of general compute power, as they move large amounts of data around for AI workflows, orchestrating across multiple agents.



VIDEO 13:59

**First look at Vera Rubin, Nvidia's next AI system that's 10 times more efficient**

"These agentic systems are spawning off different agents working as a team," CEO Jensen Huang said on Nvidia's earnings call last month. "The number of tokens that are being generated has really, really gone exponential, and so we need to inference at a much higher speed."

Huang mentioned agentic AI a dozen times on the call, and said "the best

performance-per-watt is literally everything" as hardware needs shift.

The company said in a press release that its standalone CPUs deliver significant performance-per-watt improvements in Meta's data centers.

"This is new infrastructure: Greenfield expansion of racks of CPUs whose only job is to run agentic AI," said chip analyst Ben Bajarin of Creative Strategies. "Your software is going to sit elsewhere, your accelerators are just going to run tokens, but something has to sit in the middle and orchestrate that."

## 'Quiet supply crisis'

Now, the once-sleepy central processor market is facing what The Futurum Group calls a "quiet supply crisis," predicting the CPU market growth rate could exceed GPU growth by 2028.

Leading CPU providers AMD and Intel have warned customers in China of supply shortages, according to Reuters. CPU delivery lead times are up to six months, and prices have gone up more than 10%, according to the report.

"Increases in demand are unprecedented over the last six to nine months," AMD's head of data center Forrest Norrod told CNBC in an interview.

Norrod said he doesn't see "any prospect of this slowing down or stopping anytime soon," but that AMD anticipated the lift in demand and is "working diligently" to meet it.

An Intel spokesperson told CNBC it expects inventory to hit its "lowest level" in the current quarter, "But we are addressing aggressively and expect supply improvement in Q2 through 2026."

“Wafers don’t grow on trees,” Bajarin said. “It’s not like we can just go harvest 10% more silicon wafers. There’s a crunch across the entire industry. So unfortunately, CPU wafers are constrained.”

As for whether Nvidia has seen any CPU shipment delays, Harris told CNBC, “So far, so good.”

He said Nvidia’s “robust supply chain” has been able to manage the demand, largely because many of its CPUs will be sold alongside GPUs in its rack-scale systems.



**AMD launched its 5th generation EPYC “Turin” server CPU in 2024.**

*Courtesy: AMD*

## Optimized for ‘feeding their GPUs’

Harris said Nvidia took a fundamentally different approach in design that

makes its CPUs “best suited” for data processing and agentic AI workflows, compared to the more general-purpose CPUs made by industry leaders Intel and AMD.

A big difference is in the number of cores in each CPU.

AMD’s EPYC line and Intel’s Xeon high-performance server CPUs typically have 128 cores, compared to 72 cores in Nvidia’s Grace CPU.

“If you’re a hyperscaler, you want to maximize the number of cores per CPU, and that essentially drives down the cost, the dollars per core. So that’s one business model,” Harris explained.

Instead, Nvidia designed its CPU specifically to help its star GPUs run AI workloads.

“Your single-threaded performance becomes much more important than your dollars per core because you’re trying to make sure that that very expensive resource, being the GPU, isn’t sitting there waiting,” Harris said.

Nvidia also bases its CPUs on Arm ⊞ architecture, more typically used for chips in lower-power devices like smartphones, while Intel and AMD base their CPUs on traditional x86 architecture. Introduced by Intel nearly 50 years ago, x86 is the leading instruction set that has dominated PC and server processor designs since its inception.

AMD’s Norrod said Nvidia has, “Optimized their chips very well, I think, for feeding their GPUs. They’re not well optimized for general-purpose applications.”

Indeed, Nvidia relies on more general-purpose CPUs for some of its products.

For example, Nvidia pairs its GPUs with host CPUs from Intel or AMD in its HGX Rubin NVL8 platform that customers use as the building blocks for their own AI racks.



**An Intel manufacturing technician holds an Intel Xeon 6+ data center CPU inside Intel's new Fab 52 in Chandler, Arizona in September 2025.**

*Courtesy: Intel*

## 'Platform agnostic'

Nvidia's foray into standalone CPUs comes as more of its customers are making their own Arm-based processors for their data centers.

Amazon was the first major hyperscaler to launch an in-house CPU with the release of Graviton in 2018. Google's Axion processor, released in 2024, now handles some 30% of internal applications, according to the Futurum

Group. Microsoft released its second-generation Cobalt processor in November. Arm is expected to launch its own in-house CPU this year, with Meta as an early customer.

Mercury Research estimates the server CPU market share in the last quarter of 2025 was dominated by Intel at 60%, AMD at 24.3%, and Nvidia at 6.2%, with the remaining share split among in-house Arm-based CPUs from hyperscalers like Amazon, Microsoft and Google.

In the face of insatiable need for compute, Nvidia typically takes a welcoming attitude toward competition. Keeping with that tradition, Nvidia opened up its NVLink networking technology to third-party licensing in May.

The rest of 2025 saw a flurry of NVLink deals with Intel, Qualcomm, Fujitsu, and Arm, easing the path for third-party CPUs to integrate with Nvidia GPUs in AI servers.

While these deals involve CPUs made on Arm or x86 architecture, Nvidia also now supports open instruction-set architecture RISC-V. Gaining traction in recent years, RISC-V allows companies to design custom processors without paying licensing fees to companies like Arm.

In January, Nvidia struck a deal enabling U.S. chip company SiFive to use NVLink to connect its RISC-V chip designs with Nvidia GPUs.

Harris said that no matter how the CPU demand gets filled, Nvidia's strategy remains "platform agnostic."

"We are certainly building an Arm-based CPU, but we are so invested in the x86 community, we're so invested across the ecosystem, that we're going to

have a strong position either way."

Bajarin describes Nvidia's shifting strategy as "soup-to-nuts."

"To compete, Nvidia's answer can't be you buy GPUs from us or nothing else," Bajarin said. Whether it's GPUs, CPUs or specialized hardware, "that's just the way the product has to expand to meet a diversity of workloads," he said.

Watch: CNBC's Exclusive first look at Nvidia's Vera Rubin AI system


**Read more CNBC tech news**

- Sam Altman faced 'serious questions' in meeting with lawmakers about OpenAI's defense work
- Adobe CEO Shantanu Narayen says he will step down after company installs successor
- Anthropic's Claude would 'pollute' defense supply chain: Pentagon CTO
- Palantir is still using Anthropic's Claude as Pentagon blacklist plays out, CEO Karp says



Follow your favorite stocks **CREATE FREE ACCOUNT**


**In this article**

NVDA -0.05 (-0.03%)

## MORE IN AI AGE

Choose CNBC as your preferred source on Google and never miss a moment from the most trusted name in business news.



### Elon Musk says xAI must be ‘rebuilt’ as co-founder exodus continues, SpaceX IPO awaits

Lora Kolodny



### AI agents could easily send college grad unemployment over 30%, ServiceNow CEO says

Samantha Subin



### The memory stock cycle of boom-bust-repeat is over, executives say

Kristina Partsinevelos